

Design and Analysis of Microarray Studies When Most Genes Express Differentially (or Not at All)

Naomi S. Altman
Dept. of Statistics
Bioinformatics Consulting Center
Penn State University

naomi@stat.psu.edu

A Model for Expression (One Channel)

$$Y_{gtabsqk} = \mu + G_g + T_t + B_b + S_s + A_a + Q_q + GT_{bt} + GB_{gb} + GS_{gs} + GA_{ba} + BT_{bt} + \varepsilon_{gtabsqk}$$

Y	the observed intensity of a single spot in a single channel
μ	mean intensity
G	gene effect
T	treatment effect
A	array effect
B	biological sample effect
S	subsample within biological sample
Q	probes on array for 1 gene
ε	random error

Centering and Normalization

Centering - adjustment for effects specific to one array

Normalization - adjust for "systematic" spot effects due to printing and processing.

Required because the measurement error in microarray studies is large compared to the size of effects of interest.

How Do We Center?

On array a :

Compute a measure of location (e.g.)

$$\begin{aligned}\bar{Y}_{\bullet\text{tabs}\bullet}^* &= \mu + T_t + B_b + S_s + A_a + \bar{Q} \\ &\quad + BT_{bt} + \bar{\varepsilon}\end{aligned}$$

Adjust by subtraction

$$Y_{g\text{tabs}qk} - \bar{Y}_{\bullet\text{tabs}\bullet}^*$$

How Do We Normalize?

On array a :

Compute a measure of the "systematic effect" (e.g.) by nonparametric regression:

$\tilde{Y}_{\bullet tabs(q)\bullet}^*$ a "locally" weighted average

Adjust by subtraction

$$Y_{g tabs q k} - \tilde{Y}_{\bullet tabs(q)\bullet}^*$$

The Effect of Centering and Normalization?

Centering removes an estimate of the array, biological, subsample AND treatment effect.

Normalization removes a noisier version of the array, biological, subsample AND treatment effect, as well as "local" systematic spot effects.

Normalization introduces bias due to averaging over a subset of genes, introducing an array and spot specific bias to each spot.

Why Does it Make Sense to Center and Normalize

In many studies only a small percentage of the genes are expected to differentially express in different treatments, so:

We do not expect a treatment main effect (GT is the feature of interest).

If only a small percentage of genes differentially express, GT and other gene interactions are zero for most genes, so local averaging should not introduce bias (as long as the probes are laid out at random)

But in some studies

probes are selected due to differential expression

- ESTs
- selected probes

Or, many genes differentially express (e.g.)

- comparison of organs
- transcription factor mutants

Or, many genes do not "express" (e.g.)

- mRNA samples come from immunoprecipitation studies

Using the Spiking Controls

Spiking controls are probes which do not belong to the organism under study (and should not cross-hybridize).

Matching mRNA is spiked to the samples immediately before labeling.

The mRNA can be added as a titration series which can aid in making quantitative statements linking intensity to expression levels.

What Does the Model Say about Spiking Control Expression?

For spots corresponding to the organism:

$$Y_{gtabsqk} = \mu + G_g + T_t + A_a + B_b + S_s + Q_q \\ + GA_{ba} + GB_{gb} + GT_{bt} + GS_{gs} + BT_{bt} + \\ \varepsilon_{gtabsqk}$$

For spots corresponding to spiking controls:

$$Y_{gtabsqk} = \mu + G_g + A_a + Q_q + GA_{ba} + \varepsilon_{gtabsqk}$$

Centering

Using spots from the organism:

$$\begin{aligned}\bar{Y}_{\bullet tabs \bullet}^* &= \mu + T_t + A_a + B_b + S_s + \bar{Q} \\ &\quad + BT_{bt} + \bar{\varepsilon}\end{aligned}$$

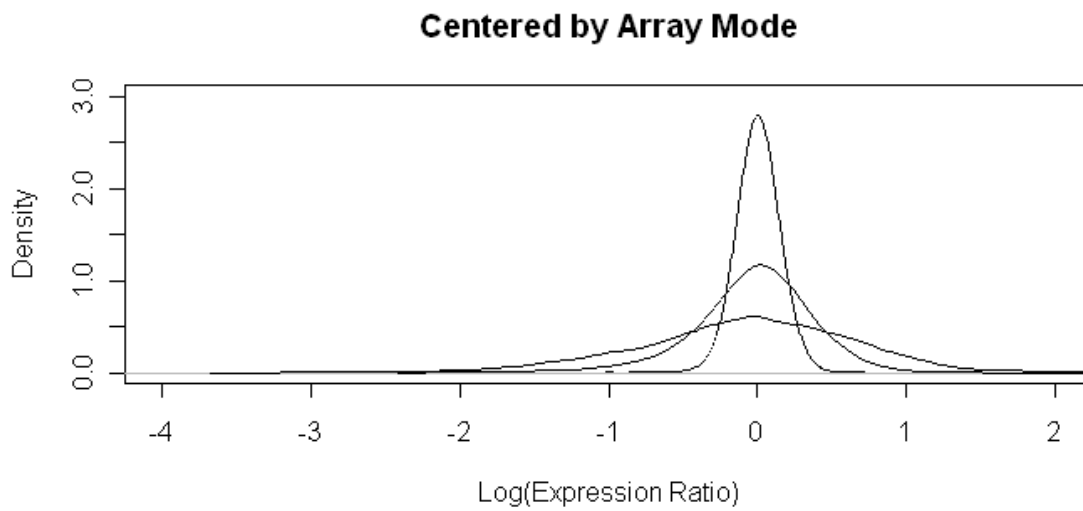
Using spiking control spots:

$$\bar{Y}_{\bullet tabs \bullet}^* = \mu + A_a + \bar{Q} + \bar{\varepsilon}$$

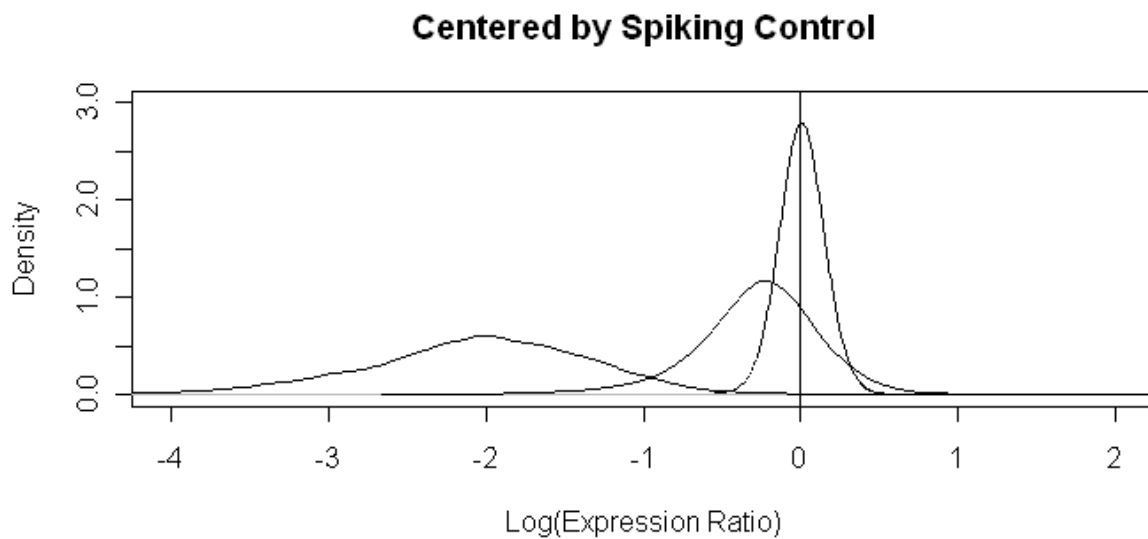
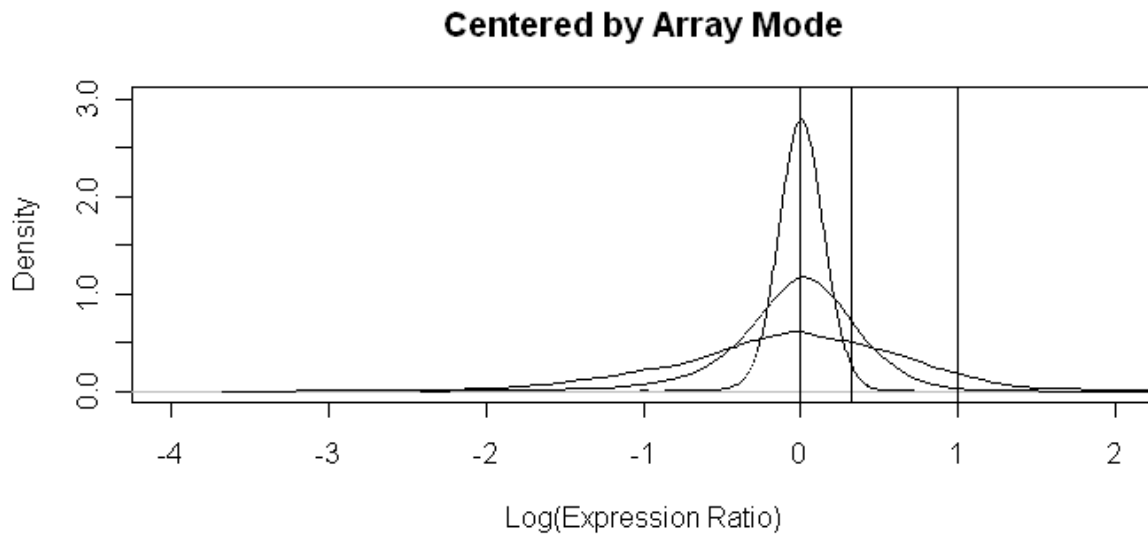
is noisier (fewer spots)

removes less "noise"

does not remove treatment effects



data on mutant yeast strains, Frank Pugh
and Kathryn Huisinga



data on mutant yeast strains, Frank Pugh
and Kathryn Huisinga

Normalization Using Spiking Controls

is not usually done because
the controls are not
sufficiently dense on the
array

How do we proceed?

In-house arrays (spotted):

- a small set of expressing genes or spiking controls needs to be spotted densely on the arrays
- spatial dispersion
- print tip dispersion
- titration series
- centering is done with spiking controls
- normalization is done with the densely spotted genes

How do we proceed?

Purchased arrays:

- we need to find a set of genes that
 - are not of interest
 - express at low level under all treatments of interest
- spike with these genes
- center using ordinary spiking controls
- normalize using the "not of interest" genes

Many thanks to

Frank Pugh's lab (Penn State)

Frank Pugh, Kathryn Huisinga, and
Andy Basehoar (centering)

Floral Genome Project (Penn State)

Claude dePamphilis, Hong Ma, Laura
Zahn (normalization)

Sarah Assman (Penn State)

(few genes expressing)

The Computational Genomics Journal
Club (Penn State)